POINT-COUNTERPOINT

On Bayesian Analysis and Hypothesis Testing in the Determination of Bioequivalence

Donald J. Schuirmann[1], Stella Grosser[1], Somesh Chattopadhyay[1] and Shein-Chung Chow[1]

Soon after the passage of the Hatch-Waxman Act, a statistical approach for determining bioequivalence (BE) was established on the basis of a frequentist hypothesis-testing approach.[1] Peck and Campbell[2] lay out concerns they have with the current method used to analyze data from comparative pharmacokinetic trials and propose a Bayesian approach to mitigate these concerns. In this article, we offer counterpoints to their argument and suggest that the current approach is statistically sound and meets statutory requirements.

Counterpoints

We focus our attention on the following topics: exchangeability and drift, statistical considerations, and specifics of the proposed Bayesian approach.

Exchangeability and drift. Peck and Campbell[2] state that generic drug approval can be "viewed qualitatively as 'Bayesian' in the sense that the newly observed BA [bioavailability] of the RLD [reference listed drug] is assumed to be 'exchangeable,' in this case, with BA-0 [BA at the time of the original approval of the RLD]." The assumption for generics, following from the 1984 amendment of the Food, Drug and Cosmetic Act,[3] known as the Hatch-Waxman Act, is only that comparable BA between the generic and the RLD implies comparable therapeutic performance between the two. The comparison necessarily uses batches of the RLD, which are being manufactured and are in the marketplace today. A patient undergoing therapy with the RLD will have been dosed with the RLD as it exists today. The batches of the RLD that were used in the pivotal human clinical trials that supported approval of the original new drug application are either long gone or long expired.

This problem has little to do with generic drugs; rather, it is a problem of maintaining the same quality and properties of the RLD over time. If BA of the RLD has changed so appreciably from "BA-0" of the clinical trial batches that the RLD is no longer safe and effective, that is a serious concern. But it is not the province of a BE study.

In the "Discussion" section of the study by Peck and Campbell,[2] one of the potential uses claimed for the proposed Bayesian approach is "evaluation of BA 'drift' of RLD...." This is a reference to their earlier discussion of the importance of BA-0. It is not clear how their analysis will accomplish this assessment. They cannot study the original clinical trial batches of the RLD in the BE study unless they intend to include some sort of historical data to address BA-0. How this, in turn, is to be done is not addressed in the article.

Distributional and other statistical considerations. In discussing the pitfalls of the current statistical framework, Peck and Campbell[2] claim that "[i]n TOST [two one-sided test] the prespecified type 1 error rate (1-sided $P < 0.05$ for each of the two one-sided hypothesis tests) is based on the untestable assumption of frequent repetition of the BE trial, and 80–90% power is required." They correctly point out that TOST assumes the subjects participating in the BE study, and the specific observations obtained from them, are a representative sample from the population of subjects who might have participated and the population of possible outcomes that those subjects might have produced. A Bayesian method would also assume this.

It is not true that "...80–90% power is required." Power is the responsibility of the entity performing the BE study.

Similarly, we agree that it is true to state "it [TOST] is merely a hypothesis test that provides no other information than whether the hypothesis is rejected or not." On the other hand, "whether the hypothesis is rejected or not" is the information we are seeking. Moreover, TOST is a size-α test,[4] a valid statistical test for average BE.

[1]Office of Biostatistics, Office of Translational Sciences/Center for Drug Evaluation and Research/US Food and Drug Administration, Silver Spring, Maryland, USA. Correspondence: Stella Grosser (stella.grosser@fda.hhs.gov)

Received 25 September 2018; accepted 30 October 2018; advance online publication January 19, 2019. doi:10.1002/cpt.1291

Another disadvantage of TOST, according to Peck and Campbell,[2] is that "it fails to provide a direct estimate of the probability of BE, much less an estimate of its entire distribution." We are left to wonder what is it, exactly, that the authors want us to do with this "entire distribution" of BE? Would they have us add additional requirements to the current requirements? This would be a clear increase in regulatory requirements (i.e., a clear increase in "the regulatory burden").

With respect to the normality assumption, additivity of the assumed statistical model is supported by pharmacokinetic models, and empirical experience supports the view that normal-theory inference methods will be valid, even with the small sample sizes of BE studies. Note, however, that (at least for standard two-period crossover studies) it is the within-subject distribution of the log-transformed end points that is assumed to be normal. The between-subject distribution may be far from normal.

Even in the Bayesian framework, some probability model is assumed for the distribution of the variables that represent the data given the prior parameters. The inference depends on correct specification of the model. The wrong Bayesian model may lead to wrong conclusions. In the case of small samples, Bayesian analysis suffers in a different way in that the prior dominates the data, and wrong prior specification may lead to wrong results.

A further threat to the normality assumption, according to Peck and Campbell,[2] is outliers.

They propose a specific Bayesian analysis, one based on the study by Kruschke,[5] which they claim is robust to the presence of outliers. The US Food and Drug Administration (FDA), however, has discouraged use of nonparametric or other robust inference procedures with standard crossover BE studies because an outlier may be telling us something important about the relative performance of the two products. A "robust" method will accommodate outliers by reducing their influence. This will permit apparently precise inference despite the presence of the outliers. That would be appropriate if the outliers represent contamination of the population of interest. But outliers may also come from a relevant subpopulation. This is the "What if my mother ..." argument (e.g., "The relative performance of the two products appears very different in subject #7 than for the rest of the subjects. What if my mother is like subject #7?") If we use a robust method, we may end up making inference only to the main population, but we will walk away from the analysis believing that we have made inference to the entire population.

A Bayesian approach. The authors propose a version of the Kruschke[5] analysis that Kruschke calls "BEST." The authors purport to propose a version of this approach, which they call "BE-BEST." However, the authors do not explain how they would modify Kruschke's[5] BEST approach, if at all. The Web link they provide takes us to a Web page purporting to implement the BEST procedure for two independent samples. How would a user implement the procedure for a crossover design?

The priors that Kruschke[5] proposes differ from the usual noninformative priors (e.g., those described by Selwyn and Hall[6]). Kruschke[5] also uses the data to determine features of the priors, which calls into question the extent to which these are truly priors. Campbell has stated previously in a different context, "The control group cannot be used [as] a source of prior information ..., especially if the objective is to show the new device is non-inferior."[7] We can replace "new device" and "non-inferiority" with "generic" and "equivalent" and the statement will apply in this context.

Furthermore, different users may have access to differing prior information, especially in a regulatory setting, leading to differing models and priors. A drawback of the Bayesian approach is that it is not completely reproducible by users and reviewers, as well as likely not to be consistent across products, while the current approach is.

Final Thoughts

For marketing approval of small-molecule generic drugs in the United States, BE assessment is not an estimation problem. It is a hypothesis-testing problem. The FDA/Center for Drug Evaluation and Research (CDER) asks whether the results of the BE study have established for each end point whether the geometric mean ratio (GMR) falls, in most cases, within 0.80 to 1.25. If the answer is yes, they are approvable. If the answer is no, they are not approvable. The FDA does not provide estimates of the GMR or other features of the *in vivo* performance of the products. Nor does the FDA permit the generic into the marketplace, with the estimates printed on the product label, thus enabling each individual user or prescribing physician to make his or her own decision as to whether to use the generic. In the US system, the generic is approved with an AB rating, or else it is not approved. This is a two-decision problem with two kinds of errors, exactly the kind of problem that hypothesis testing addresses, rather than a problem of point or interval estimation.

Over the years, the FDA/CDER has considered ways to take into account other aspects of the relative performance of the generic and the RLD, besides the overall GMR and has issued guidance to this effect in particular cases. The authors' proposed BE-BEST might provide information on other aspects of the relative performance of the two products, once more details of the procedure are established. It would then be up to the CDER to decide what use, if any, to make of that additional information.

ACKNOWLEDGMENTS

The authors would like to thank Thomas Permutt and two anonymous referees, whose comments substantially improved this article.

FUNDING

No funding was received for this work.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

DISCLAIMER

This article reflects the views of the authors and should not be construed to represent the US Food and Drug Administration's views or policy.

1. Schuirmann, D.J. A comparison of the two-one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **15**, 657–680 (1987).
2. Peck, C.C. & Campbell, G. Bayesian approach to establish bioequivalence: how and why? *Clin. Pharmacol. Ther.* <https://doi.org/10.1002/cpt.1288>.

3. US Government Publishing Office. Drug Price Competition and Patent Term Restoration Act of 1984. Public Law 98-117. <https://www.gpo.gov/fdsys/pkg/STATUTE-98/pdf/STATUTE-98-Pg1585.pdf>. Accessed September 22, 2018.
4. Chow, S.C. & Shao, J. A note on statistical methods for assessing therapeutic equivalence. *Control. Clin. Trials* **23**, 515–520 (2002).
5. Kruschke, J.K. Bayesian estimation supersedes the t test. *J. Exp. Psychol. Gen.* **142**, 573–603 (2013).
6. Selwyn, M.R. & Hall, N.R. On Bayesian methods for bioequivalence. *Biometrics* **40**, 1103–1108 (1984).
7. Campbell, G. "Bayesian statistics at the FDA: the trailblazing experience with medical devices" presented at Rutgers Biostatistics Day, April 3rd, 2009 <http://www.stat.rutgers.edu/iob/bioconf09/slides/Campbell.pdf>. Accessed October 22, 2018.